March 2004

Inferential Risk Analysis in Support of Standards for Emissions of Hazardous Air Pollutants from Hazardous Waste Combustors

Final Report

Prepared by

RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

Prepared for

Office of Solid Waste
U.S. Environmental Protection Agency
401 M Street SW (5307W)
Washington, DC 20460

EPA Contract Number 68-W-98-085

RTI Project Number 08860.000.002

[This page intentionally left blank.]

Table of Contents

Section **Page**

List of Figures

List of Tables

[This page intentionally left blank.]

1.0 Background and Purpose

1.1 Background

The U.S. Environmental Protection Agency (EPA) is developing maximum achievable control technology (MACT) standards for hazardous waste combustors (HWCs). Phase I of the HWC MACT rulemaking was promulgated on September 30, 1999, for three categories of HWCs: incinerators, cement kilns and lightweight aggregate kilns (64 FR 52828). These standards were promulgated under the joint authority of the Clean Air Act (CAA) and the Resource Conservation and Recovery Act (RCRA). As part of the Phase I rulemaking, EPA conducted an extensive assessment of risks to human health and the environment from hazardous air pollutants (HAPs) emitted by Phase I sources.

In Phase II of the HWC MACT rulemaking, EPA plans to develop standards for industrial and commercial/institutional boilers (which encompass both liquid and solid fuel boilers) and hydrochloric acid production furnaces.[1] The technology-based MACT standards developed under Phase II are intended to supersede the emission standards for these sources established pursuant to RCRA and codified at 40 CFR Part 266, Subpart H. However, Section 3004(a) and 3004(q) of RCRA require EPA to develop standards that are protective of human health and the environment. For this reason, EPA is conducting a comparative analysis of the risks from hazardous-waste-burning boilers and industrial furnaces (as compared to risks from incinerators from Phase I) as part of the Phase II rulemaking.

On July 24, 2001, the United States Court of Appeals for the District of Columbia Circuit (the Court) granted the Sierra Club's petition for review and vacated portions of the Phase I rule (Cement Kiln Recycling Coalition v. EPA, Docket No. 99-1457a, July 24, 2001).

On October 19, 2001, EPA and the petitioners that challenged the Phase I emission standards filed a joint motion asking the Court to stay the issuance of its mandate for four months to allow EPA time to develop interim standards (a request that was subsequently granted by the Court). In the joint motion, EPA agreed to take several actions. For instance, EPA agreed to issue final replacement standards that fully comply with the Court's mandate by June 14, 2005.[2] Importantly, EPA also agreed to issue standards for Phase II sources concurrently with the Phase I replacement rule.

[1] Industrial and commercial/institutional boilers include process heaters that meet the RCRA definition of a boiler.

[2] EPA also committed to (1) publish an interim rule with revised emission standards, and (2) finalize several compliance and implementation amendments to the Phase I rule. These interim standards and compliance and implementation amendments were promulgated on February 13 and 14, 2002 (67 FR 6792 and 67 FR 6968).

Table 1-1 gives the emission standard levels for the currently proposed MACT standards for HWCs. EPA considered a number of options in the development of the proposed standards. The emissions that are projected to occur under each of those options were the subject of this study.

Table 1-1. Proposed Standards for Existing Sources

	Incinerators	**Cement Kilns**	**Lightweight Aggregate Kilns**	**Solid Fuel-Fired Boilers[a]**	**Liquid Fuel-Fired Boilers[a]**	**Hydrochloric Acid Production Furnaces[a]**
Dioxin/ Furans (ng TEQ/dscm)	0.28 for dry APCD and WHB sources;[f] 0.40 for others	0.20 or 0.40 + 400°F at APCD inlet	0.40	CO or THC standard as a surrogate	0.40 for dry APCD sources; CO or THC standard as surrogate for others	0.40
Mercury	130 ug/dscm	64 ug/dscm[b]	67 ug/dscm[b]	10 ug/dscm	3.7E-6 lb/MMBtu[b,e]	Total chlorine standard as surrogate
Particulate Matter	0.015 gr/dscf [h]	0.028 gr/dscf	0.025 gr/dscf	0.030 gr/dscf [h]	0.032 gr/dscf [h]	Total chlorine standard as surrogate
Semivolatile Metals (lead + cadmium)	59 ug/dscm	4.0E-4 lbs/MMBtu[e]	3.1E-4 lb/MMBtu[e] and 250 ug/dscm[c]	170 ug/dscm	1.1E-5 lb/MMBtu[b,e]	Total chlorine standard as surrogate
Low Volatile Metals (arsenic + beryllium + chromium)	84 ug/dscm	1.4E-5 lbs/MMBtu[e]	9.5E-5 lb/MMBtu[e] and 110 ug/dscm[c]	210 ug/dscm	1.1E-4 lbMMBtu[d,e]	Total chlorine standard as surrogate
Total Chlorine (hydrogen chloride + chlorine gas)	1.5 ppmv[g]	110 ppmv[g]	600 ppmv[g]	440 ppmv[g]	2.5E-2 lb/MMBtu[e, g]	14 ppmv or 99.9927% system removal efficiency

[a] Particulate matter, semivolatile metal, low volatile, and total chlorine standards apply to major sources only for solid fuel-fired boilers, liquid fuel-fired boilers, and hydrochloric acid production furnaces.

[b] Standard is based on normal emissions data.

[c] Sources must comply with both the thermal emissions and emission concentration standards.

[d] Low volatile metal standard for liquid fuel-fired boilers is for chromium only. Arsenic and beryllium are not included in the low volatile metal total for liquid fuel-fired boilers.

[e] Standards are expressed as mass of pollutant contributed by hazardous waste per million Btu contributed by the hazardous waste.

[f] APCD denotes “air pollution control device,” WHB denotes “waste heat boiler.”

[g] Sources may elect to comply with site-specific, risk-based emission limits for hydrogen chloride and chlorine gas.

[h] Sources may elect to comply with an alternative to the particulate matter standard.

1.2 Purpose and Scope

The overall purpose of this study was to provide risk analysis support for the Phase I replacement (hereafter referred to as "New Phase I") and Phase II HWC MACT rulemaking. Specifically, the objective was to develop as necessary and compare risk-affecting variables (emissions, stack characteristics, meteorological data, and population data) from Phase I sources for which human health risk estimates are available (hereafter referred to as "Old Phase I") to these same variables from the New Phase I/Phase II sources to determine the extent to which the replacement/new sources might be judged similar to the Old Phase I sources. If the Old Phase I risk assessment determined that risks were acceptable for a certain chemical for a certain Phase I source category (e.g., commercial incinerators), and the New Phase I/Phase II risk-affecting variables for a specific category (e.g., Phase II liquid boilers) were found to be similar to the Old Phase I variables, then, all other things being equal, one could reasonably infer that the (unknown) human health risks from the New Phase I or Phase II source category for this chemical would not be expected to differ significantly from the Old Phase I risks.

There are many differences between an Old Phase I and New Phase I or Phase II source category that may affect relative risks between the two. Emission rates, stack characteristics, meteorological variables, population differences (number and spatial distribution), population makeup, local environmental conditions, and land use are all important factors that may affect risks. To the extent that any of these factors differ between the source categories being compared, the risks may be different. Because it is impossible as a practical matter to thoroughly compare all these factors, those factors believed to be most important and for which data were either available or could be developed within the resource constraints of this study were selected for comparison. The factors for which data were readily available were chemical-specific emission rates, stack characteristics, and some meteorological variables. These available data were then augmented by collecting additional meteorological data and collecting and statistically analyzing site population data. Thus, emission rates, stack characteristics, meteorological variables, and population characteristics are the four overall characteristics that were believed to most significantly affect risks. Any remaining risk-affecting factors (e.g., topography, soil characteristics, surface water characteristics, land use) were not compared, and subsequent relative risk inferences should be understood to implicitly assume that there are no significant differences among these factors between Old Phase I combustor sites and New Phase I/Phase II sites or that these factors are relatively insignificant contributors to risks. This should be recognized as a significant limitation, particularly for categories of New Phase I/ Phase II sources having a small number of sites where unaccounted for site-specific factors could be important contributors to risks.

The Old Phase I risk assessment provides the risk benchmarks from which relative risk inferences are made based on the results of the comparative analyses performed in this study. The following overview is excerpted from the Old Phase I risk assessment technical background document prepared by Research Triangle Institute (RTI, 1999).

1.3 Old Phase I Risk Assessment

On April 19, 1996, EPA proposed rules to revise standards for hazardous waste combustors, which include hazardous-waste-burning incinerators, cement kilns, and lightweight aggregate kilns. The rule was proposed under joint authority of the CAA, as amended, and RCRA, as amended. HWCs emit HAPs that are listed under Section 112(d) of the CAA. EPA proposed national emission standards for hazardous air pollutants (NESHAP) pursuant to Section 112(d) of the CAA that establish emission standards based on application of maximum achievable control technology. Hence, these standards are referred to as MACT standards.

These MACT standards are technology-based standards; they are not risk-based. These facilities, however, are also covered by RCRA Sections 3004(a) and 3004(q), which require EPA to develop standards that are protective of human health and the environment. To meet the current MACT requirements under the CAA and to satisfy RCRA's requirement, the Phase I [Old Phase I] risk analysis was conducted to support the MACT standard rulemaking for HWCs. EPA's express intent was to minimize duplication in regulations and regulatory actions. Accordingly, the MACT standards for incinerators, cement kilns, and lightweight aggregate kilns were developed under CAA authority. Consideration of human health and ecological risk allowed EPA to satisfy the requirements of both RCRA and the CAA.

The risk assessment conducted for the final rule covered the same source categories evaluated in the April 19, 1996, proposed rules: incinerators, cement kilns, and lightweight aggregate kilns. For the risk assessment for the final rule (since vacated), three subcategories were added for incinerators: commercial incinerators, onsite incinerators (small), and onsite incinerators (large). Waste heat boilers, which are associated with some incinerators, were evaluated separately.

The [old] Phase I risk assessment was a multimedia, multipathway assessment that addressed direct exposures to constituents released to the atmosphere by HWC units and indirect exposures due to movement of constituents into the food chain. The risk assessment addressed both human health risks (cancer effects and noncancer effects) as well as ecotoxicological risks. Constituents assessed were 7 congeners of chlorinated dioxin and 10 congeners of chlorinated furan, 3 species of mercury, the 11 metals that were modeled for the proposed rule (antimony, chromium VI, chromium III, arsenic, lead, barium, nickel, beryllium, selenium, cadmium, silver, and thallium), 3 additional metals (cobalt, copper, and manganese), particulate matter (PM), hydrochloric acid, and chlorine gas.[3]

Individual risk for receptors that could be enumerated using U.S. Census and Census of Agriculture data (for most types of cancer and noncancer effects) was characterized through the use of cumulative risk distributions, which were constructed by weighting sector-level individual risk estimates by the number of individuals located in that sector and then pooling those weighted risk estimates.[4] These pooled risk estimates were then ranked according to risk magnitude, and specific percentiles of interest were identified. These percentiles can be interpreted as representing the risk level experienced by the individual located at that point on

[3] It should be noted that silver, cobalt, copper, and manganese were not included in the present study due to limited data on emissions and/or low toxicity.

[4] Individual risks for receptors that could not be enumerated, such as persons engaged in subsistence farming or subsistence fishing, were also characterized. This was done through the use of cumulative distributions that were constructed from sector-level individual risk estimates, which were equally weighted with respect to the numbers of individuals.

the risk distribution (i.e., central tendency or high-end risk estimates can be identified). These cumulative risk distributions included a number of factors designed to make them representative of the receptors for which they were developed:

- They reflected the location and density of receptors across study areas. Study areas encompassed the area surrounding a facility out to a radius of 20 kilometers.
- They were based on central tendency exposure parameters (key exposure pathways were subject to exposure parameter variability analyses designed to incorporate this additional source of variability into the characterization of risk, e.g., exposures to dioxins and furans in home-produced beef and milk and methyl mercury in recreationally caught freshwater fish).
- They were based on a 16-sector template, which enhanced resolution in assessing exposure. The 16 sectors were the quadrants defined by the intersection of radii to the north, east, south, and west and rings at 2, 5, 10, and 20 kilometers.

1.3.1 Phase I Risk Assessment Overview

The following provides an expanded discussion of the 1999 risk assessment. For more details, the reader is referred to the risk assessment background documents for the 1999 final rule (RTI, 1999).

The risk assessment for the 1999 rule was based on an analysis of 76 facilities. This sample of facilities comprised 66 facilities selected by stratified random sampling and an additional 10 facilities that had previously been selected in the risk analysis for the 1996 proposed rule. This sample of facilities represented nearly half of the facilities known to be burning hazardous waste at that time in cement kilns, lightweight aggregate kilns, and incinerators within the continental United States.[5] Emissions were estimated for each facility based on site-specific stack gas emission concentrations and flow rates measured during trial burns or compliance tests.[6] EPA used a design level for projecting what a facility’s emissions would be when in compliance with the MACT standards: facilities emitting below the design level during trial burn and compliance tests were assumed to continue to emit at the levels measured in the tests while facilities emitting above the design level were assumed to reduce their emissions to the design level. The design level was taken as 70 percent of the MACT standard.[7] The percentage reduction in emissions required to meet the design level was applied to each chemical constituent to which the standard applied.

[5] The 1999 risk assessment did not include solid or liquid fuel-fired boilers or hydrochloric acid production furnaces burning hazardous waste, since these sources were not within the scope of the rule.

[6] For facilities where stack gas measurements were not available, data were imputed by random selection from a pool of measurements for similar units.

[7] This reflects the reality that facilities cannot emit continuously at the level of the MACT standard and simultaneously be in compliance with the standard at all times, due to variability in facility emissions. This approach is consistent with the assumption made in the cost and economic analysis for the rule that facilities emitting below the design level would not need to retrofit with new control technologies.

The analysis for the 1999 rule assessed the risks to the entire population of individuals living within 20 kilometers (or about 12 miles) of the sample of facilities. A study area composed of 16 sectors was established for each facility.[8] U.S. Census and Census of Agriculture data were used to estimate the numbers and ages of individuals living in farm households by type of farm and the population of individuals living in non-farm households for each of the 16 sectors. Individuals were grouped into four primary age categories: 0 to 5 years, 6 to 11 years, 12 to 19 years, and 20 years of age and older. Additional age categories were used for assessing risk from PM. Within each study area, three or four bodies of water were chosen for analysis based on their proximity to the sample facility and the likelihood of their being used for recreational purposes or if they were known to supply drinking water to the surrounding community. The watershed of each waterbody was delineated out to a distance of 20 kilometers from the facility. The analysis assumed that a portion of the households in each study area would engage in recreational fishing, based on the prevalence of recreational fishing in national surveys, and that recreational anglers would fish at all of the waterbodies delineated in a given study area.

The analysis for the 1999 rule assessed risks from multiple exposure pathways, including inhalation; incidental ingestion of soil; consumption of drinking water; consumption of home-produced fruits and vegetables; and consumption of home-produced meat, milk, poultry, fish, and eggs. Exposure pathways varied depending on the particular receptor population (e.g., home gardeners, dairy farmers, recreational anglers) and the types of activities that lead to human exposures. Age-specific rates of mean daily food intake and media contact rates were used in conjunction with sector-specific concentrations of chemical contaminants in media and food to calculate the total dose to an individual from all exposure pathways combined. Distributions of individual risks were generated for each receptor population by weighting sector estimates of individual risk with sector-specific population weights and facility-specific sampling weights. Lifetime average daily dose was used for assessing cancer risk, and average daily dose (reflecting less than lifetime exposure) was used for assessing risks of noncancer effects. For certain exposure pathways (e.g., ingestion of dioxins in beef and milk and mercury in fish), an exposure parameter variability analysis was performed. This was accomplished using a combination of exposure factor distributions (i.e., age-specific distributions of food intake rates and duration of exposure), a life table analysis (to adjust for changes in age-specific intake rates over the duration of exposure), and Monte Carlo sampling (to generate distributions of risks for a receptor population as a whole from sector-specific risk distributions).

The risk assessment for the 1999 rule also included a screening-level ecological analysis. The analysis compared model-estimated media concentrations to media-specific ecotoxicological criteria that are protective of multiple ecological receptors. These included criteria for soils, surface waters, and sediments. For assessing ecological risks from dioxins in surface water, direct comparisons were made of estimated intakes by fish-eating birds and mammals of 2,3,7,8-tetrachlorodibenzo(p)dioxin (TCDD) toxicity equivalents (TEQ) to ecotoxicological benchmarks for TCDD in order to account for the widely differing rates of bioaccumulation of the various TCDD and dibenzofuran congeners in fish.

[8] The sectors were defined by the intersection of concentric rings at 2, 5, 10, and 20 kilometers and radii extending to the north, south, east, and west.

1.3.2 Phase I Risk Assessment Results Summary

Detailed risk results for all human receptor populations and for the ecological risk analysis are contained in a separate six-volume series of documents (RTI, 1999). The risk results for human receptor populations are summarized here in terms of the "margin of exposure." In general, the margin of exposure, or MOE, is the ratio of the benchmark dose or point of departure (POD), the reference dose (RfD), or the reference concentration (RfC) to the estimated dose or air concentration for a given receptor. In analogous fashion, the MOE is taken here to be the multiple by which the estimated exposure is below an exposure level that could be a cause for concern from a risk management perspective, in this case a lifetime excess cancer risk of 1E-5 and a hazard quotient (HQ) = 1 for noncancer effects. The higher the MOE, the greater the margin between the exposure and a risk level of concern.[9]

Tables 1-2, 1-3, and 1-4 show the MOE for the chemical contaminants evaluated in the 1999 risk assessment for the Phase I source categories for the regulatory baseline and assuming compliance with the 1999 MACT standards. Exposure factors (e.g., intake rates and duration of exposure) were tailored to the specific receptor populations evaluated in the 1999 assessment. The MOE listed in the table for a given percentile is the lowest for any route of exposure (inhalation or ingestion) for the receptor population having the highest estimated risks (whether cancer or noncancer).[10, 11] For dioxin TEQ, MOEs are given for the cancer slope factors from both EPA's 1985 health assessment document and from EPA's recent draft dioxin reassessment.[12, 13] With few exceptions, all MOEs are rounded down and reported to one significant figure.

[9] An MOE less than 1 (shown in bold in Tables 1-1, 1-2, and 1-3) indicates the estimated exposure exceeds a level of concern. In this instance, the multiple by which the estimated exposure is above the level of concern would be represented by the inverse of the MOE.

[10] The MOE values are based on the 90th percent upper confidence limits on the risk estimates whenever confidence limits, which account for sampling error, were available.

[11] The MOE values do not account for emissions from other units at hazardous waste combustion facilities, nor do they account for background exposures from other, non-hazardous waste sources. Background exposures can be significant, particularly for dioxin, mercury, and lead.

[12] See U.S. EPA (1985) and U.S. EPA (2000). Note that the latter document is a draft document and does not necessarily represent EPA policy regarding the characterization of risks from dioxins and related compounds.

[13] In addition, the MOE for chlorine was adjusted to account for changes in the RfC since the 1999 risk assessment (from 1 $\mu g/m^3$ to 0.2 $\mu g/m^3$). See U.S. EPA (1999).

Table 1-2. 1999 MACT Rule—MOE for Cement Kilns[a]

Constituent	50th Percentile	90th Percentile	95th Percentile	99th Percentile
Baseline				
antimony	1,000,000	1,000	1,000	-
arsenic	20,000	1,000	1,000	500
barium	50,000	2,000	1,000	1,000
beryllium	1,000,000	20,000	20,000	10,000
cadmium	3,000	500	300	100
chlorine	300	20	20	10
chromium (III)	20,000,000	5,000,000	2,000,000	1,000,000
chromium (VI)	10,000	10,000	10,000	3,000
dioxin TEQ (1985)	50	5	3	1.4
dioxin TEQ (2000)	8	**0.8**	**0.5**	**0.2**
hydrogen chloride	1,000	300	200	100
lead	50	30	20	20
mercury	50	3	2	1.3
nickel	100,000	10,000	5,000	3,000
selenium	10,000	2,000	1,000	1,000
thallium	10,000	500	100	30
MACT Standards				
antimony	1,000,000	1,000	1,000	-
arsenic	20,000	2,000	1,000	1,000
barium	50,000	2,000	1,000	1,000
beryllium	1,000,000	30,000	20,000	10,000
cadmium	10,000	2,000	1,000	500
chlorine	400	40	20	20
chromium (III)	20,000,000	5,000,000	2,000,000	1,000,000
chromium (VI)	10,000	10,000	10,000	5,000
dioxin TEQ (1985)	100	10	5	2
dioxin TEQ (2000)	10	1.7	**0.8**	**0.3**
hydrogen chloride	1,000	300	300	200
lead	1,000	500	500	300
mercury	50	5	3	1.7
nickel	100,000	10,000	5,000	3,000
selenium	10,000	2,000	1,000	1,000
thallium	10,000	500	100	50

[a] MOE below an exposure associated with a cancer risk of 1E-5 and an HQ of 1.
Dash (-) indicates percentile not estimated due to small sample size or an insufficient spread of modeled risks.

Table 1-3. 1999 MACT Rule—MOE for Lightweight Aggregate Kilns[a]

Constituent	50th Percentile	90th Percentile	95th Percentile	99th Percentile
Baseline				
antimony	100,000	10,000	10,000	10,000
arsenic	10,000	3,000	2,000	1,000
barium	500,000	20,000	20,000	10,000
beryllium	300,000	50,000	50,000	20,000
cadmium	20,000	2,000	1,000	1,000
chlorine	1,000	100	60	20
chromium (III)	10,000,000	3,000,000	2,000,000	1,000,000
chromium (VI)	30,000	3,000	2,000	1,000
dioxin TEQ (1985)	20	3	1.4	**0.5**
dioxin TEQ (2000)	4	**0.4**	**0.2**	**0.08**
hydrogen chloride	100	50	30	20
lead	1,000	500	500	300
mercury	500	50	20	-
nickel	50,000	3,000	1,000	1,000
selenium	500,000	300,000	100,000	100,000
thallium	100,000	30,000	20,000	5,000
MACT Standards				
antimony	100,000	10,000	10,000	10,000
arsenic	10,000	5,000	5,000	2,000
barium	500,000	30,000	20,000	10,000
beryllium	300,000	100,000	50,000	30,000
cadmium	30,000	5,000	5,000	2,000
chlorine	1,000	200	60	60
chromium (III)	20,000,000	10,000,000	3,000,000	2,000,000
chromium (VI)	50,000	5,000	5,000	1,000
dioxin TEQ (1985)	100	50	20	10
dioxin TEQ (2000)	10	8	3	2
hydrogen chloride	500	100	100	50
lead	1,000	1,000	1,000	1,000
mercury	500	50	30	-
nickel	50,000	3,000	1,000	1,000
selenium	500,000	300,000	100,000	100,000
thallium	100,000	30,000	20,000	5,000

[a] MOE below an exposure associated with a cancer risk of 1E-5 and an HQ of 1.
Dash (-) indicates percentile not estimated due to small sample size or an insufficient spread of modeled risks.

Table 1-4. 1999 MACT Rule—MOE for Incinerators[a]

Constituent	50th Percentile	90th Percentile	95th Percentile	99th Percentile
Baseline				
antimony	10,000	50	50	-
arsenic	10,000	100	100	50
barium	500,000	10,000	5,000	5,000
beryllium	1,000,000	10,000	10,000	3,000
cadmium	20,000	500	500	20
chlorine	200	20	20	5
chromium (III)	50,000,000	3,000,000	2,000,000	1,000,000
chromium (VI)	10,000	500	100	20
dioxin TEQ (1985)	300	10	5	1.0
dioxin TEQ (2000)	50	1.7	**0.8**	**0.17**
hydrogen chloride	10,000	1,000	200	100
lead	20	10	10	8
mercury	50,000	300	100	50
nickel	100,000	20,000	10,000	1,000
selenium	3,000,000	30,000	20,000	10,000
thallium	100,000	10,000	1,000	100
MACT Standards				
antimony	50,000	300	300	100
arsenic	50,000	500	500	500
barium	2,000,000	50,000	30,000	10,000
beryllium	1,000,000	20,000	10,000	10,000
cadmium	100,000	5,000	2,500	500
chlorine	400	200	60	10
chromium (III)	100,000,000	10,000,000	10,000,000	5,000,000
chromium (VI)	20,000	2,000	1,000	200
dioxin TEQ (1985)	1,000	50	20	10
dioxin TEQ (2000)	100	8	4	1.7
hydrogen chloride	10,000	1,000	1,000	500
lead	1,000	500	500	300
mercury	50,000	300	100	50
nickel	200,000	30,000	10,000	5,000
selenium	3,000,000	30,000	20,000	10,000
thallium	1,000,000	30,000	2,000	1,000

[a] MOE below an exposure associated with a cancer risk of 1E-5 and an HQ of 1.
Dash (-) indicates percentile not estimated due to small sample size or an insufficient spread of modeled risks.

1.4 Overview of Comparative Analysis

The objective of the comparative analysis is to evaluate the differences between the sources subject to the 1999 HWC MACT rule and the sources subject to the currently proposed MACT rule with respect to four key factors. These factors may be described generally as emission rates, stack gas characteristics, meteorological conditions, and exposed populations. Each factor (or megavariable) comprises one or more subvariables (e.g., stack height) or frequency variables (e.g., frequency of wind speeds of 1 meter per second or less). For each variable, formal hypothesis tests were performed to determine if the variable was significantly different between the sources being compared. For example, to compare the projected emissions of a pollutant under the currently proposed rule and the emissions under the 1999 MACT rule, a hypothesis test was performed on the upper tail (e.g., 90th percentile emission rate) of the distribution of emissions as they are characterized for the proposed rule versus the distribution of emissions as they were characterized for the 1999 MACT rule, to determine if they were significantly different. For the source categories that were not a part of the previous risk assessment (i.e., liquid fuel-fired boilers, solid fuel-fired boilers, and hydrochloric acid production furnaces), comparisons were made with the incinerator source category from the 1999 rule. For the other source categories, comparisons were made with their counterparts from the 1999 rule. The comparisons were conducted assuming compliance with the 1999 MACT standards and the proposed MACT standards (with standards at either the floor or beyond-the-floor levels, as the case may be).

Similar to the risk assessment done for the 1999 rule, emissions for the proposed rule were estimated for each facility based on site-specific stack gas concentrations and flow rates measured during trial burn or compliance tests.[14] The emission projections assume that sources will design their systems to meet an emission level below the proposed standards in order to be in compliance with the standards at all times. This level is called the “design level” and accounts for the expected variability in emissions during facility operations. Where the test data indicate the emissions are below the design level, it was assumed the source would continue to emit at the levels measured in test. For sources emitting above the design level, it was assumed the source would reduce emissions to the design level.[15]

The comparative analysis tested the upper tail (90th or 75th percentile), lower tail (10th or 25th percentiles), or median of the distributions depending on the particular variable and its relationship to risk. For emissions, higher emission rates lead to higher risk and, therefore, hypothesis tests were performed on the upper tail of the distribution. For stack height, the lower tail was tested because lower stack heights lead to higher risks to individuals living in the vicinity of the source. Hypothesis tests were also performed on a second stack gas characteristic, termed the buoyancy flux. Buoyancy flux represents the heat released to the atmosphere through

[14] Although emissions and stack data were imputed by random selection from a pool of measurements for similar units for sources where stack gas measurements were not available, these data were not used in the statistical comparisons.

[15] In the 1999 rule, the design level was taken as 70 percent of the standard. For the currently proposed standards, the design level is generally the lower of (1) 70 percent of the standard; or (2) the arithmetic average of the emissions of the best performing sources, as defined by the CAA.

the stack and, as such, is determined by the temperature and volume of the stack gases. The greater the heat release, the greater the buoyancy of the stack gas plume and the lower the impact on nearby, ground-level receptors. Because lower buoyancy is associated with higher risk, hypothesis tests were performed on the lower tail of the distribution. If hypothesis testing could not be performed at the desired percentile (e.g., 90th / 10th percentile) due to data limitations, the next closest percentile was used (e.g., the 75th / 25th percentile or the 50th percentile). For variables that affect risk but do not have the same relation to risk under all conditions (or at least a wide range of conditions), the median of the distribution was tested. Most of the meteorological variables fall into this category.

To simplify the analysis, all comparisons were done at the site level (i.e., by facility). For sites with multiple stacks, emissions were totaled across stacks and stack characteristics (i.e., stack height and buoyancy flux) were averaged. Meteorological data were taken from the closest available meteorological station.

Population data were subjected to a regression analysis prior to hypothesis testing. The purpose of the regression analysis was to analyze the distribution of population with respect to distance from the site. Specifically, an exponential regression model was fitted to the sector-level population data for farm and non-farm households for each hazardous waste combustion facility site where such data had been collected and analyzed. Population data originally collected for the risk analysis for the 1999 rule was supplemented with data collected for the additional categories of sources subject to the currently proposed rule (i.e., solid and liquid fuel-fired boilers and hydrochloric acid production furnaces). The regression analysis provided two parameters for each site, alpha and beta. The alpha parameter is the intercept and the beta parameter is the slope of the log-scale transformation of the regression model. Alpha is a measure of the population density adjusted for the frequency of wind direction in each quadrant and, as such, is an indicator of the "time-exposed" population (i.e., the time the wind blows towards the population). Beta is a measure of the distribution of population with distance from a source. For a uniform population density about a source, a beta coefficient equal to 2 is expected because area increases as the square of the distance. For betas less than 2, the population is distributed relatively closer to the source and for betas greater than 2, the population is distributed relatively further from the source. A lower beta is associated with higher risk. Therefore, for the purpose of the comparative analysis, hypothesis tests were performed on the lower tail of the distribution of betas across the sites. By contrast, alpha is more closely associated with overall risk to the population (e.g., population risk) and, therefore, hypothesis tests were performed on the median of the alpha distribution.

Once the hypothesis tests were performed, scores were assigned to the outcomes. If the hypothesis test indicated no significant difference (at a p value of 0.1), a neutral score (0) was assigned. If the hypothesis test did indicate a statistically significant difference ($p = 0.1$), a positive score (+1) or negative score (-1) was assigned, depending on whether the observed difference would be expected to lead to lower risk (+1, a "risk-favorable" outcome) or higher risk (-1, a "risk-unfavorable" outcome).[16] These risk inferences were made with the assumption

[16] For example, if the 90th percentile emission rate for liquid fuel-fired boilers is less than 90th percentile emission rate for incinerators (as characterized in the 1999 rule) and the difference is statistically significant, a score of +1 would be assigned because lower emissions lead to lower risk, all other things being the same.

of "all other things being equal." In instances where the differences in percentile values are statistically significant yet the effect on risk is ambiguous, an "888" score (a "not zero" outcome) was assigned. Where hypothesis testing could not be performed due to data limitations, a "999" score (insufficient data) was assigned. These scores are termed percentile scores because they are direct comparisons of the relevant percentiles of the distribution (10th / 90th, 25th / 75th, or 50th percentile).

Because all other things are not equal, hypothesis testing was also performed on correlations between variables. If the differences in the correlation coefficients were not statistically significant ($p = 0.1$), a neutral score (0) was assigned. If the difference was statistically significant, a positive score (+1) or negative score (-1) was assigned, depending on whether the observed difference would be expected to lead to lower risk (+1) or higher risk (-1).[17] These scores are termed correlation scores because they reflect correlations among the various parameters. All combinations of variables were tested for differences in their correlation coefficients.

A weight-of-evidence approach was used for assessing the overall direction of risk relative to the 1999 rule. This was accomplished by aggregating and weighting the scores from the individual comparisons. First, an aggregate correlation score was formed from the set of correlation scores for a given variable by giving equal weights to the correlations with all the other variables. Then, an aggregate score for the variable was formed by equally weighting the aggregate correlation score and the percentile score for that variable.[18] Then, the aggregate score for each frequency variable associated with a subvariable and each subvariable associated with a megavariable were given equal weights and aggregated. Finally, the megavariable factors (i.e., emissions, stack characteristics, meteorology, and population) were weighted and aggregated. (The weights applied to the megavariables were based on an evaluation of the performance of the comparative analysis methodology, as described in Section 6.0.) This process was applied to the subset of comparisons where the only possible scores were "+1," "-1," or "0" (i.e., where the outcome of the hypothesis test was considered to be unambiguously risk-favorable, risk-unfavorable, or neutral) and a "Grand Score" was computed. The Grand Score can range from -1 (least favorable) to +1 (most favorable).

In addition, reliability indices were assigned to each of the hypothesis tests and aggregated in similar fashion to provide a "reliability index." The reliability index reflects the amount of data available to perform the hypothesis tests and, therefore, the confidence in the test results. The reliability index can range from 1 (most reliable) to 4 (least reliable). The grand score was then divided by the associated reliability index to generate a "normalized" Grand Score. In addition, the number of +1, -1, 0, 888, and 999 scores were tallied, weighted (as just described for the Grand Score but using the full set of comparisons), and normalized by the number of tests performed. The resulting counts represent the weighted fraction of tests with

[17] For example, a higher correlation of emissions with population (as represented by the population density parameter, alpha) would be assigned a score of -1 because higher emissions in more populated areas leads to higher risk. Conversely, a lower correlation of emissions and population would be assigned a score of +1 because lower emissions in more populated areas (and higher emissions in less populated areas) leads to lower risk.

[18] Here, "variable" refers to a variable, subvariable, or frequency variable.

each type of outcome (i.e., weighted by correlation vs. percentile and frequency variable, subvariable, and megavariable). This information, together with a specific set a decision rules, enabled conclusions to be made regarding the anticipated impact on risk vis-a-vis the 1999 risk assessment. The information generated by comparative analysis is intended to assist EPA in making judgments as to whether the emission standards it is proposing for HWCs are generally protective of human health and the environment, as required under RCRA.

2.0 Phase I and II Source Categories

For the purposes of this analysis, subsets of the Old Phase I source categories were used as the basis for comparison to New Phase I/Phase II categories. The categories for Phase I (Old and New) and Phase II HWCs are defined and described below.

2.1 Phase I Categories

The three Phase I categories are all incinerators, cement kilns (CKs), and lightweight aggregate kilns (LWAKs).

- **All Incinerators**
  - *Onsite Incinerators* function as part of a larger commercial manufacturing operation and handle hazardous wastes generated specifically by that operation (these facilities do not burn wastes for other companies for profit). Because onsite facilities play a support role and are not dependent on earning profit through hazardous waste combustion, they are often smaller than commercial facilities (their size is dependent on the type of operation they support) and burn a limited variety of wastes.
  - *Commercial Incinerators* function specifically as commercial facilities that earn revenue by burning hazardous wastes. As such, the incinerators in this category are often larger (i.e., larger throughput) and burn a greater variety of wastes than those in the onsite category.
  - *Waste Heat Boilers* (WHB) recover excess heat generated in the incineration process as a thermal source for industrial applications rather than releasing it directly to the environment. A subset of both onsite and commercial incinerators have waste heat boilers.
- **Cement Kilns (CKs)** are the pyroprocessing step in chemically combining a variety of raw mineral materials into cement. The kilns are rotary kilns (long, cylindrical, slightly inclined furnaces) into which raw materials are fed at the upper end and product removed at the lower end. Cement production is an energy intensive process using coal, oil, and natural gas as fuel. More recently constructed plants tend to be more fuel efficient, but hazardous waste is used as a replacement fuel to reduce the operating costs of the kilns. Some waste selectivity is practiced as elements present in the waste can make the end product unsuitable for certain uses.

- **Lightweight Aggregate Kilns (LWAKs)** are the pyroprocessing step in generating a coarse aggregate used in the production of lightweight concrete products, such as concrete block, structural concrete, and pavement. The process works by expanding the raw materials, such as clay, shale, slate, and blast furnace slag, to about twice their original volume, typically in a rotary kiln. This process is energy intensive, using coal, coke, fuel oil, and natural gas as fuel. In some plants hazardous waste is used to replace these more expensive fuel types to reduce operating costs.

2.2 Phase II Categories

The two Phase II categories are boilers and halogen acid furnaces (HAFs).

- **Boilers** are an enclosed device using controlled flame combustion for recovering and exporting thermal energy in the form of steam, heated fluid, or heated gases for onsite process needs. These units must, by definition, maintain a thermal energy recovery of at least 60 percent and export and utilize at least 75 percent of the recovered energy. Boilers can be further disaggregated based on fuel source type into solid boilers (SBs) and liquid boilers (LBs). A liquid fuel-fired boiler is a device that meets the definition of a boiler and burns any combination of liquid and gas fuels, but no solids. A solid fuel-fired boiler is a device that meets the definition of a boiler and burns solid fuels, including pulverized or stoker coal.

- **Halogen Acid Furnaces (HAFs)** are an integral component of a chemical production facility that processes hazardous waste with a minimum as-generated halogen content of 20 percent to produce an acid product with a minimum halogen content of 3 percent. These acid products are subsequently used in a manufacturing process.

In addition to these distinctions, incinerators (all types) and LBs can be further categorized as “dry” or “not dry.” “Dry” refers to a dry air pollution control system (e.g., electrostatic precipitators or baghouse). “Not dry” can be either a wet air pollution control device (APCD) (e.g., wet scrubber) or no back-end controls. The distinction is made in the context of dioxins, because dioxin formation is increased in dry air pollution control systems. Dioxin formation is also increased in incinerators that have WHBs and therefore, such incinerators are included in the “dry” category, regardless of the type of back-end control.

Appendix A provides a complete list of the Phase I and II combustors by category type, HWC phase to which they belong, and stack versus site ID. In some cases, multiple individual stacks occur at the same facility site. The data presented in Appendix A are the full population (“universe”) of stacks/facilities (i.e., these are not samples from a larger population). These data also include an indicator flag denoting those Old Phase I facilities, out of the universe of Old Phase I facilities, that were randomly sampled (stratified sampling) for the original Phase I risk assessment (RTI, 1999). A total of 76 unique Old Phase I site-level facilities were sampled for the original Phase I risk assessment.

A summary of the number of unique HWC site-level facilities (i.e., accounting for multiple stacks at a single site) by Phase and category is given in Table 2-1.

Table 2-1. Number of Unique Sites by Combustor Phase and Category

Phase and Category	Number of Sites
Old Phase I All Incinerators	144
Old Phase I CK	18
Old Phase I LWAK	5
Phase II LB	55[a]
Phase II Dry LB	6
Phase II Not Dry LB	50
Phase II SB	4
Phase II HAF	8
New Phase I All Incinerators	77
New Phase I Dry All Incinerators	18
New Phase I Not Dry All Incinerators	57
New Phase I CK	14
New Phase I LWAK	3

[a] Although the sum of the “dry” plus “not dry” sites does not necessarily sum to the total number of sites (see earlier explanation of these categories), ideally they would not sum to more than the total number of sites. Indeed, they do for Phase II LBs (by 1) because two sites (LA008213191 and TXD008092793) have multiple stacks including one in each category. Thus, these sites have been included in both categories.

[This page intentionally left blank.]

3.0 Variables Considered and Data Compared

As mentioned previously, a complete list of all risk-affecting variables would include emission rates, stack characteristics, meteorological variables, population differences (number and spatial distribution), population makeup, local environmental conditions, and land use. Because it is impossible as a practical matter to thoroughly compare all these factors, those factors believed to be most important and for which data were either available or could be developed within the resource constraints of this study were selected for comparison. Those selected data used for the subsequent comparative analyses are described in this section. They fall into four general categories:

- Chemical-specific emission rates
- Stack characteristics
- Meteorological conditions
- Population characteristics.

For purposes of the subsequent comparative analysis of data distributions, it is useful to think of these data distributions as comprising up to three categories, termed here megavariables, subvariables, and frequency variables. A megavariable is an overall category of data (i.e., the four general categories listed above). A subvariable is a more specific category within a megavariable (e.g., stack height for the stack characteristics megavariable or windspeed within the meteorological megavariable). Finally, a frequency variable is relevant for data that describe time frequencies for certain conditions (e.g., percent of time that the windspeed is less than a specific velocity). The data distributions used in the subsequent comparative analyses are described below in this hierarchical context.

3.1 Emission Rates

Chemical-specific emission rate is obviously an important risk-affecting variable: as emission rates increase, risks increase if all other factors remain the same. For each subcategory of Old Phase I combustor emissions (e.g., Old Phase I All Incinerators) compared to a subcategory of Phase II or New Phase I (e.g., New Phase I CK), a distribution of emission rates (kg/yr) was developed by individual chemical of interest. These chemical-specific emission rates reflect aggregate emissions at a site (i.e., where multiple stacks occur at any site, emissions are summed over all stacks).[1] With the exception of sites where multiple stacks occur, site-level emission rates used in the distributions are measured-only data. Emissions data that were not

[1] Site-level emissions were aggregated on a source category basis (i.e., incinerators and boilers at the same site were not combined for purposes of the analysis).

actually measured, but rather imputed from other data, were not used (to avoid artificially increasing the number of "observations" by using imputed values even though the sample size has not really changed), with one exception: where multiple stacks occur at a site, and some of these stacks have measured emissions rates and others have imputed rates, the imputed rates were used in estimating the aggregate site emission rates. It was assumed that the potential for introducing biases in aggregate emission rates by using imputed data in these multistack situations was preferable to either ignoring the imputed data and using only the measured data (which would underestimate site-level total emissions) or not including these sites at all in the data distribution. In addition, in a number of cases the units were sufficiently similar to tested units (so-called "sister" units) that testing was not deemed necessary for purposes of permitting or compliance. Therefore, imputation of data from a sister unit for use in estimating site-level emissions was considered thoroughly appropriate.

3.2 Stack Characteristics

Characteristics of combustor stacks are also important risk-affecting factors. Two stack characteristic subvariables are considered—height and buoyancy flux. High stacks induce more atmospheric mixing and dispersion of chemicals than do low stacks. Therefore, all other factors being the same, high stacks reduce risks relative to low stacks. Buoyancy flux is a measure of the heat contained in the exit gases and therefore, is an indicator of the potential rise of the plume (and therefore, atmospheric dilution) due to buoyancy effects arising from the difference in the density of the exit gases and the surrounding air. Therefore, all other factors being the same, risks decrease with increasing buoyancy flux. Buoyancy fluxes were calculated as part of this study for each stack as a function of stack gas exit velocity, stack diameter, exit gas temperature, and long-term average ambient site temperature as follows (Briggs, 1975):

$$F_b = g v_s d_s^2 \left(\frac{\Delta T}{4 T_s} \right) \tag{3-1}$$

where

F_b = buoyancy flux (m^4/s^3)
g = gravitational acceleration (m/s^2)
v_s = stack gas exit velocity (m/s)
d_s = stack diameter (m)
ΔT = Ts - Ta (K)
Ts = stack gas temperature (K)
Ta = ambient temperature (K).

Where multiple stacks occur at a site, average stack height and average buoyancy flux were calculated and used in the subvariable distributions. Only actually measured values were used in the distributions. In instances where multiple stacks occur at a site and some of these stacks have imputed data, still only the measured data were used to estimate the site average. In general, stack data were reported for all stacks for which emissions test data were available. (Unlike aggregate emission rates, where omitting stacks causes an obvious bias in the site total, using measured-only data to estimate site averages was felt to be preferable to using imputed data. For buoyancy flux, if any of the input variables used in the buoyancy flux calculation were imputed, the buoyancy flux was also considered imputed and not used.

3.3 Meteorological Conditions

Five meteorological subvariables were identified as presumed important risk-affecting factors: atmospheric mixing height, windspeed, atmospheric stability class, wind direction variability, and mean precipitation.

Mixing height is the height above the ground surface through which relatively vigorous vertical mixing occurs. Mixing height was estimated using the interpolation scheme employed in the RAMMET meteorological processor, which uses the twice-daily mixing heights from the nearest National Weather Service upper air observation site, coupled with the stability category determined for the hour. The subvariable mixing height is comprised of three frequency variables that give the percentage of time that the site mixing height is less than or equal to 500 meters, 1000 meters, and 1500 meters, respectively.

Wind speed affects risks by creating turbulence and dispersion of contaminants, thereby reducing risks. However, the turbulence created by the wind also reduces plume rise, which increases risk. Therefore, although wind speed is an important determinant of risk, it is not possible to draw a general conclusion about what effect it will have because it depends on the values of other parameters, such as stack height and buoyancy flux. The windspeed subvariable is comprised of four frequency variables that give the percentage of time that the site windspeed is less than or equal to 1 m/s, 3 m/s, 5 m/s, or 10 m/s, respectively.

Atmospheric stability is a classification scheme that attempts to take into account both the effects of mechanical turbulence and the effects of thermal turbulence, or convection. Unstable conditions promote greater atmospheric mixing and contaminant dispersion. However, greater mixing can bring an elevated plume to the ground more quickly, particularly in the presence of convection. Therefore, although atmospheric stability is an important determinant of risk, it is not possible to draw a general conclusion about what effect it will have because it depends on the values of other parameters, such as stack height and buoyancy flux. The stability class subvariable is comprised of two frequency variables that give the percentage of time that the site stability class is within (1) Class A, B, or C (least stable categories) or (2) Class E, F, or G (most stable categories).

Wind direction variability affects risks inversely. If a site has a predominant wind direction (low variability), then that condition increases risks for the population downwind. A higher variability does not necessarily decrease short-term air concentrations, but it does mitigate high-end risks by not exposing the same population repeatedly (and therefore, having the effect of reducing long-term air concentrations). The wind direction variability subvariable is represented by the “circular variance” (Mardia, 1972) of wind direction. The circular variance was calculated for this analysis for each site given that site’s wind direction frequency data. These data describe the percentage of time that the wind direction is in each of 16 radial directions (i.e., north, north-northeast, northeast, east-northeast, east, etc.). The circular variance is normalized on a scale of 0 to 1. Values approaching 1 reflect widely dispersed wind direction, while values approaching 0 reflect a strong predominant wind direction. Therefore, all other factors being equal, a site having a higher circular variance than another site would expect decreased exposure and risk.[2] The circular variance (S_0) was calculated (Mardia, 1972) as

$$S_0 = 1 - \overline{R} \qquad (3\text{-}2)$$

$$\overline{R} = (\overline{C}^2 + \overline{S}^2)^{1/2} \qquad (3\text{-}3)$$

$$\overline{C} = \sum_i F\Theta_i Cos(\Theta_i) \qquad (3\text{-}4)$$

$$\overline{S} = \sum_i F\Theta_i Sin\left(\Theta_i\right) \qquad (3\text{-}5)$$

where

Θ_i = angle from due north of ith wind direction (0 degrees for i=1, 22.5 for i=2, etc.)
$F\Theta_i$ = relative frequency that wind direction is toward Θ_i.

Meteorological data used in the comparative analyses are observations on the full universe of Old Phase I, Phase II, and New Phase I facilities. The meteorological data were generated using the meteorological preprocessor PCRAMMET (U.S. EPA, 1995). The preprocessor pairs hourly surface observations with upper-air measurements. For each meteorological station modeled, five years of surface and upper-air data were used.

Hourly surface meteorological data used in air dispersion modeling were processed from the Solar and Meteorological Surface Observation Network (SAMSON) CD-ROM (U.S. DOC and U.S. DOE, 1993). The variables include temperature, pressure, wind direction, windspeed, opaque cloud cover, ceiling height, current weather, and hourly precipitation. Twice daily mixing height data were gathered from the Radiosonde Data of North America CD-ROM (NCDC, 1997).

All meteorological data were considered observations for the purpose of the present analysis.

[2] We can conclude this as a general matter except in the relatively unlikely circumstance that population is distributed in the same way as the distribution of wind direction.

3.4 Population Characteristics

Population characteristics are associated with human health risk in three important ways that were considered in this analysis. First, the mere existence of human receptors in the vicinity of HWCs introduces the possibility of exposure and risk. Therefore, the first population characteristic subvariable is related to the total number of receptors within the vicinity of each site. More people leads to more total exposure and therefore, greater risk. We used a radial distance of 20 kilometers to represent the exposed population because that was the distance used to characterize the exposed population in the original Phase I risk assessment. The total population-related subvariable is denoted as "alpha" (for reasons that will be described shortly) and is a measure of the average density of the receptor population within 20 kilometers of each combustor site. Therefore, all other factors being the same, the higher the value of alpha, the higher the risk.

Secondly, the spatial distribution of receptors within this 20 kilometer radius also affects risks. It is presumed in this analysis that the closer receptors are located to the HWC, the higher is their exposure and risk. This simply reflects the fact that, in general, air concentrations are greater closer to the source, before they have had time to be dispersed and diluted. This is not always true, of course. For example, high stacks with relatively distant population centers may give rise to higher risks than having those same receptors located closer to the source, because the exhaust plume may travel some considerable distance before dropping to ground level. Nonetheless, for the sources being analyzed, the highest concentrations are expected to occur in the first kilometer or two, unless there is significant elevated terrain further from the source.

As described in Appendix B, the spatial distribution of receptor-specific population within the 20 kilometer radial distance around each combustor site was estimated as a continuous parameter, "beta," from geographic information systems (GIS) data. The GIS data consist of census estimates of the number of receptors located within 16 sectors around each combustor site. The 16 sectors are defined by radial distances from the site (2 km, 5 km, 10 km, and 20 km) and the four cardinal directions (north, east, south, west) from each combustor site. These data were then used to fit regression models of the general form

$$P_{ij} = \alpha D_{ij}^{\beta} \tag{3-6}$$

where

P_{ij} = the number of exposed receptors[3] at site *i* within radial distance *j*
D_{ij} = radial distances
α = regression constant (to be estimated for each site)
β = regression constant (to be estimated for each site).

[3] The population data were obtained by site, radial distance, and sector. The exposed population at any radial distance j was estimated as the total population at j weighted by the percent of time that the wind is in the direction of each of the four sectors, i.e., NE, SE, SW, and NW.

Alpha is a measure of the average receptor density at the site while beta is a parameter that quantifies the spatial distribution of these receptors. If the population is approximately uniformly distributed with land area, beta would take on values near 2.0, because land area increases with the square of the radial distance from the site (just as the area of a circle increases with the square of its radius); hence, the number of receptors would increase with the square of distance. Values of beta less than 2.0 imply that receptors are more "source-concentrated" (i.e., have higher densities close to the source than farther away). For example, a beta of 1.0 implies that receptor numbers are increasing linearly with radial distance from the source. Given that land area (at least total land area if not residential land use) necessarily increases as the square of distance from the source, it can be seen that population density must be higher close to the source. This situation bodes unfavorably for risk, all other factors being equal. Conversely, values of beta greater than 2.0 imply that population densities increase with radial distance, a risk-favorable condition.

Finally, in addition to total site population density (alpha) and site spatial distribution (beta), risk is associated with certain subpopulations of receptors. For example, exposure and cancer risk to dioxins was found to be highest to children of dairy farmers who consume home-produced milk. Accordingly, the population GIS data were collected specific to certain subpopulations of receptors and the regression model (Equation 3-1) was fit separately for each of these subpopulations, so that the estimated constants alpha and beta are available on a subpopulation-specific basis. The subpopulations considered that were used in the comparative analysis (other subpopulations were also analyzed, as described in Appendix B) were

- Child residents aged 0 to 5 years (for lead)
- Farm families (for dioxin)
- Total population (all other chemicals).

As described in Appendix B, the Old Phase I, Phase II, and New Phase I variable distributions for the alpha and beta parameters do not reflect the universes of these facilities, but rather are random samples. For Old and New Phase I, the distributions reflect the 76 sampled stacks for the original Phase I risk assessment. For Phase II, 41 sites were selected from a precursor ("Old Phase II universe") that varied slightly from the "Phase II universe" (current) as used elsewhere in this report.[4] Of these 41 sites,[5] all solid boilers and HAFs in the (old) Phase II universe are included in the sample. For the liquid boilers (59 sites in the old Phase II universe), 29 of these were randomly sampled for development of the alpha and beta parameters. The 29 samples out of 59 sites ensured that the probability of selecting a site that would be in the upper 10th percentile of the 59-site risk distribution would exceed 90 percent. (This is the same criterion that was used to select the sites that were analyzed in the risk assessment for the Old phase I universe.) The value of 29 was determined by solving the following equation

[4] Of 70 sites in the "old Phase II" universe, 5 of these sites are not in the (new) Phase II universe. Conversely, of 67 sites in the (new) Phase II universe, 2 of these sites are not in the old Phase II universe.

[5] Appendix B shows population data for 40 sites, not 41. It is noted that one of these Appendix B sites (TXD008092793) includes 2 HAF stacks and 1 LB stack. Thus, it counts as a HAF "site" (with both stacks) as well as a LB "site" for purposes of this analysis.

(developed by RTI) for *n* using a statistical significance level (α, not to be confused with the population “alpha”) of 0.95 and percentile (p) of 90, and rounding up to the nearest integer:

$$n \geq \frac{\ln(\alpha)}{\ln\left(p/100\right)} \tag{3-7}$$

For new Phase I sources, population data for 35 of the 77 incinerator sites, 11 of the 14 cement kiln sites, and 3 of the 3 lightweight aggregate kiln sites were available from the original Phase I risk assessment and were used to develop the alpha and beta distributions.

The complete distribution of site samples for determination of population characteristics across the New Phase II sites and combustor categories is included in the Appendix A data, under the field “New Phase II Population Parameter Sampled?” Appendix A also provides the distribution of Old Phase I stacks and combustor categories sampled out of the Old Phase I universe under the field “Old Phase I Risk-Modeled?”

[This page intentionally left blank.]

4.0 Statistical and Graphical Methods

4.1 Principles of Interval Estimation and Hypothesis Testing

4.1.1 Overview

In this section, basic principles of confidence interval estimation and hypothesis testing will be reviewed. Interval estimation of various Phase I and II population parameters (e.g., means and percentiles of emissions) will be described in Section 4.2, while tests for hypotheses about these parameters (such as “Phase I and II mean emission rates are equal”) will be presented in Section 4.3. In this section, for simplicity, the term “Phase II” should be understood to refer generically to either Phase II or New Phase I data. In Section 4.4, a graphical method (cumulative density [CDF] plots) for comparing Phase I and Phase II distributions will be explained. Finally, special problems associated with inferences based on tests or interval estimates computed from small samples will be discussed in Section 4.5. It is noted that SAS analyses of comparisons between Phase I and II data were performed using the methods described in this section, and all methods are reviewed here for completeness. Not all of the results of these methods, however, were considered appropriate to carry forward into the final comparative analyses, implemented by the computer program RelRisk and described in Section 6. The methods used by RelRisk are described in Section 6.

4.1.2 Precision and Bias

Population parameters (e.g., Phase I mean emission rates) are estimated by selecting a random sample from the population of interest (e.g., Phase I incinerators), making measurements on the sample members, and then generating an appropriate statistic (e.g., the sample mean emission rate). It is intuitive that not every sample estimate is good, in the sense that its value is close to that of the unknown population parameter. In particular, the sample statistic may be biased, imprecise, or both. Figure 4-1 illustrates four possibilities as distinct shot patterns around a bull’s-eye; each bull’s-eye represents conceptually a population target parameter, while the shots are statistical estimates of the parameter obtained from repeated sampling of the target population. Figure 4-1A shows a tight clustering of the shots outside of the bull’s-eye and thus illustrates the case where the statistical estimates are biased but precise. Such bias—or systematic errors—may result from various causes, including measurement bias and sampling bias. Measurement bias occurs when the instrument is inaccurate (e.g., from improper or inadequate calibration, from the effect of interferences in a chemical analysis). Sampling bias can occur when the sampling mechanism is not random and favors some members of the population over others. Note that increasing the sample size does not affect the bias; hence, to avoid bias, one must be sure to use proper instrumentation and sample designs.

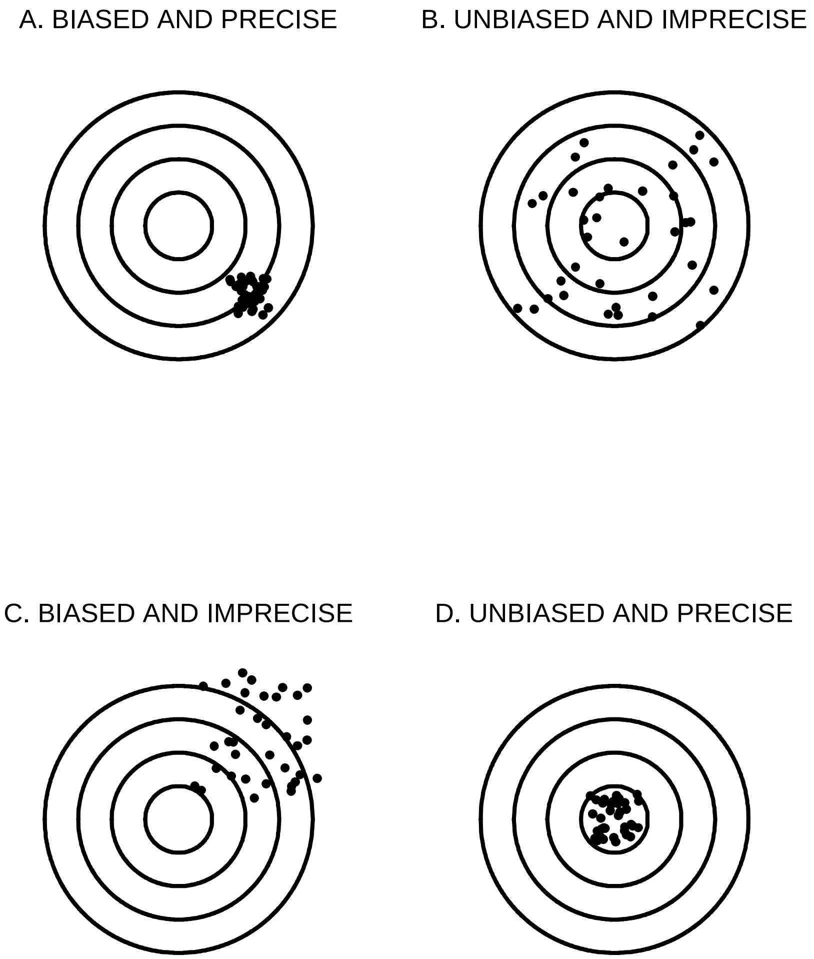

Figure 4-1. Bias and precision as represented by shot patterns on a target.

The statistical estimates in Figures 4-1B and 4-1D are not skewed in any particular direction, thus there does not appear to be any bias. However the pattern in 4-1B is highly dispersed around the true parameter value. This is indicative of considerable heterogeneity in the target population and/or considerable random measurement error (e.g., from an imprecise instrument), which leads to imprecision in the sample statistics. Imprecision and heterogeneity are reflected in increased dispersion of the statistical estimates around the value of the target parameter. In situations like Figure 4-1B, a larger sample size could be used to reduce the uncertainty in the estimates; the result would be something like the tight clustering shown in Figure 4-1D.

Environmental quality decisions frequently are made on the basis of probabilities derived from the sample estimates by the application of statistical inferential procedures. The computed probabilities support the acceptance or rejection of two competing hypotheses, the null and the alternative. For example, in a study comparing the emission rates of Phase I and II combustors, one might consider the null hypothesis (H_0) to be that the emission rates of the Phase I and II populations are approximately equal, while the alternative hypothesis (H_a) is that they are different from one another. Given these two competing hypotheses, imprecision and/or bias inherent in the sample estimates of the desired population parameters create the potential for two types of decision errors. (Note: Since the bias is unknown, one must generally assume it to be zero in making calculations.)

So-called Type I errors occur when the null hypothesis is incorrectly rejected, e.g., Phase I emissions that are approximately equal to Phase II emissions are erroneously judged to be significantly different. A Type II error occurs when the emission rates of the Phase I and II combustors are erroneously judged to be approximately equal. Type I and II errors are often compared to the judicial errors of convicting an innocent man (Type I) or of letting a guilty man go free (Type II). When we decrease the probability of one error, we coincidentally (and nearly inevitably) increase the probability of the other. Thus a subjective decision is usually made to guard against one at the expense of the other. Much like the judicial system analogy, the scientific community has focused on protecting against Type I errors rather than Type II errors. There are two statistical tools available to help scientists and regulators make decisions, based on sample estimates, such that the error rates are taken into account: confidence intervals and hypothesis tests.

4.1.3 Quantifying Sampling Error: Confidence Intervals

Population variability and sampling and measurement error obviously affect the imprecision in sample estimates. The imprecision results in "sampling error" in the estimates computed from the sample. The standard error of a sample estimate of a population parameter provides a quantitative expression of the sampling error. One way to account for sampling error is to use the sample estimate and its standard error to construct confidence intervals for the population parameter. Such an interval has some known probability (e.g., a confidence level of 95 percent) of containing the true population parameter, in the sense that if the entire sampling and measurement process were repeated a large number of times, then the percentage of the intervals that actually cover the true value would equal the prescribed confidence level. The confidence interval is therefore a statement about the confidence we have in the sample estimate of a population parameter, θ.

Algebraically, the general expression for a two-sided 1-α ×100% confidence interval is

$$\Pr\left[a_1 \leq \Theta \leq a_2\right] = 1 - \alpha \tag{4-1}$$

where

$1-\alpha$ = the desired confidence level
a_1 = the $1-\alpha/2$ lower bound of the confidence interval for θ
a_2 = the $1+\alpha/2$ upper bound of the confidence interval for θ.

The bounds a_1 and a_2 of the confidence interval are the "confidence limits"; their magnitudes are a function of the estimate and its standard error and depend on the sample size and the probability distribution of the estimate. By convention, the confidence level, 1-α, is expressed as a percent (e.g., 95 percent). If we specify α=0.05, then we are in effect saying that intervals we have constructed have a 5 percent chance of not covering the true population parameter and a 95 percent chance of covering it.

Two-sided 100×(1-α)% confidence intervals are appropriate when one desires a sample estimate of an unknown population parameter together with a measure of the amount of uncertainty in the estimate. Also, the 100×(1-α)% confidence intervals of two estimates (e.g., the sample mean lead emission rates of Phases I and II) can be compared to determine the degree of overlap. The width of the confidence interval (i.e., a_2-a_1) provides a measure of the precision of the estimate; the smaller the width, the greater the precision. The confidence interval width depends on the size of the standard error of the estimate (a function of the variability of the measurement and the sample size), the sample size itself, and the specified confidence level (α). When each of the other two factors is held constant, the following changes will result in narrower confidence intervals for population parameters such as emission rates:

- Decreasing the variance
- Decreasing the confidence level (e.g., going from 95 percent to 80 percent confidence)
- Increasing the sample size.

In general, the variance is comprised of both inherent variability within the population and the measurement error uncertainty, and only the latter is subject to control (e.g., by using a more precise instrument). Specific formulae, and details for the construction of confidence intervals for a variety of population parameters, are presented in Section 4.2.

4.1.4 Quantifying Sampling Error: Hypothesis Tests

In the decision-making process, hypothesis testing provides an alternative to the comparison of confidence intervals for accounting for the uncertainty in sample data. There are several possible sources of uncertainty, including

- Sampling variation specific to the design employed to collect the data
- Intrinsic natural variation among population members
- Temporal or spatial variation
- Measurement or laboratory error
- Model misspecification error (e.g., in Monte Carlo risk assessments).

In an environmental study to compare combustor emissions in two populations (e.g., Phase I incinerators vs. Phase II liquid boilers), one may have an *a priori* hypothesis that emission rates are greater in one population than in another. In that case, either of two pairs of one-sided null and alternative hypotheses may be evaluated,

$$H_0\!:\theta_1 \le \theta_2 \quad vs. \quad H_a\!:\theta_1 > \theta_2 \qquad (4\text{-}2)$$

or

$$H_0\!:\theta_1 \ge \theta_2 \quad vs. \quad H_a\!:\theta_1 < \theta_2 \qquad (4\text{-}3)$$

where (for example)

θ_1 = the Phase I incinerator population emission rate
θ_2 = the Phase II liquid boiler population emission rate.

Equation 4-2 evaluates an upper one-sided alternative hypothesis, while Equation 4-3 evaluates the complementary lower one-sided alternative.

Alternatively, if there is no strong *a priori* evidence to support the direction of the difference, a more general two-sided alternative hypothesis may be specified:

$$H_0 : \theta_1 = \theta_2 \quad vs. \quad H_a : \theta_1 \neq \theta_2 \tag{4-4}$$

Having selected the population parameters of interest (e.g., the mean TEQ emission rate) and chosen null and alternative hypotheses appropriate to the decision under consideration, the next step is to choose a statistical test that can be used to determine which hypothesis (H_0 or H_a) is better supported by the sample data. Statistical tests are mathematical models that are used to predict the distributions of test statistics when the null hypothesis is true. These are the distributions one would expect to obtain from conducting thousands of surveys or experiments and plotting the frequencies of the computed test statistics. These distributions, called sampling distributions, reflect the uncertainty in the sample estimates of the values of the test statistic.

Statistical tests are commonly named for their sampling distributions (e.g., t-test or chi-squared test). The test statistics themselves are usually simple algebraic functions of the sample statistics. For example, the test statistic for the test against either the two-sided or the one-sided alternatives for comparing the mean lead emissions of Phase I ($\overline{X}_1$) to the mean of Phase II ($\overline{X}_2$) is a function of the sample size (*m* and *n*), the means ($\overline{X}_i$), and the variances (S_i^2):

$$\frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{1/m + 1/n}} \tag{4-5}$$

where
m = Phase I sample size
n = Phase II sample size

$$S_p = \sqrt{\left[(m-1)S_1^2 + (n-1)S_2^2\right] / (m+n-2)}. \tag{4-6}$$

Statistical theory ensures that when certain assumptions hold and the null hypothesis is true, the sampling distribution of the test statistic shown in Equation 4-5 will be a t-distribution with m+n-2 degrees of freedom. Thus the associated statistical test is called a t-test. There is actually an entire family of t-distributions, each with a different degrees of freedom.

The sampling distributions of the test statistics should not be confused with the population distributions from which the samples have been collected. Whereas the latter are the distributions of the combustor emissions under study, the former are statistical models of the behavior of statistics calculated from samples of the combustor sites. Some statistical tests, called parametric tests (e.g., t-tests), require that the natural populations be normally distributed, while others, called nonparametric tests, make no assumptions about the distribution of the natural populations.

All statistical hypothesis tests are mathematical models. In the case of t-tests and chi-square tests, the distribution of the test statistic computed from the sample data is modeled as (respectively) a t-distribution or a chi-square distribution. Like all models, the validity of the predicted distributions depends on assumptions made about the underlying processes that are being modeled. In the case of two-sample t-tests for the population mean, the following assumptions are made:

- The variable being analyzed is a continuous random variable that is normally distributed in the two target populations.
- The sampling units used to compute the sample means from which the t-statistic (Equation 4-4) was calculated were independently distributed in their respective target populations (i.e., there is no temporal or spatial autocorrelation among the sampling units).
- There was no systematic error associated with measuring the response on the sampling units (e.g., spectrophotometers or laboratory assays were correctly calibrated and applied).
- The null hypothesis is true.

The distribution of the test statistic under the null hypothesis is the basis for determining whether the data support the null hypothesis or the alternative. For example, if we have a sample of 15 combustors in each phase, the expected distribution of t-statistics under the null hypothesis is a t-distribution with Degrees of freedom = (15+15-2) = 28. Ninety-five percent of the t-statistic values in such a distribution will lie between -2.048 and +2.048. Suppose we measure lead emissions for the thirty facilities and compute the t-statistic for lead emissions using Equation 4-5. If the absolute value of the t-statistic computed from our sample were 2.5, then we would conclude that there is less than a 5 percent chance that our sample came from such a t-distribution. This suggests that one or more of the above four assumptions is *not* true. If we have previously verified the first three assumptions, we can conclude that our sample *does not* support the null hypothesis. If we have not verified the assumptions, we cannot draw any conclusions from the t-test. Rejection of the null hypothesis provides evidence in favor of the alternative that the Phase I and Phase II lead emission rates are significantly different from one another. Due to the assumption required by the t-test that the variables be normally distributed, the comparisons made in the present study rely on other test statistics, such as the chi-squared and Wilcoxon tests.

The probability used as the cutoff for accepting or rejecting the null hypothesis is called the significance level. By declaring a significance level of 5 percent, we are saying that even though there is a 5 percent probability that a t-statistic with an absolute value $\geq$ 2.048 could have come from the t-distribution associated with the null-hypothesis, this probability is so small that we believe it is more reasonable to think that the data actually came from populations with different mean lead emission rates. Thus the significance level is just the Type I error rate (α) that the investigator has decided, *a priori*, that he or she is willing to tolerate.

4.2 Estimates of Population Parameters from Sample Data

4.2.1 Overview

Statistical comparisons between Phase I and II data were of two types: (1) direct comparisons of selected distributional parameters, and (2) comparisons of correlation coefficients between selected pairs of Phase I variables with correlation coefficients for the same variable pairs in the Phase II data. "Comparison" means implementation of statistical hypothesis tests for testing equality and/or comparing the overlap of confidence intervals.[1] Interval estimates were computed for distributional parameters of the variables of interest, including the mean, median (50th percentile), variance, and, as permitted by data limitations, the 10th, 25th, 75th, and 90th percentiles. In this section, the sample point and interval estimators of these population parameters are described.

In this section and in Section 4.3, "X" denotes a variable (e.g., stack height) that is being compared between a combustor subgroup in Phase I (e.g., incinerators) and a subgroup in Phase II (e.g., liquid boilers). The units of analysis were site-level means or totals of these variables, computed by aggregating over the stacks of a particular subgroup at a given site. For example, the mean TEQ emission rate of the liquid boiler subgroup was a computed as the grand mean of the means (taken over stacks, within each site) of TEQ emissions from each of the sites at which liquid boilers were operating during the study. The equations that follow provide estimates of various parameters (e.g., means, medians, variances) of Phase I and Phase II combustor subpopulations. For convenience, the estimated parameters are not subscripted in any way. Thus, for example, $X_{(p)}$ will be used to represent the pth population of variable X percentile in either a Phase I or a Phase II subgroup.

4.2.2 Interval Estimates of the Mean of a Normal Population

When the values of X are approximately normally distributed in the subgroup population, the100×1-α% confidence interval for the population mean, μ_x, is (Snedecor and Cochran, 1980)

$$\bar{x} - t_{1-\alpha/2,n-1}\sqrt{\frac{s_x^2}{n}} \quad to \quad \bar{x} + t_{1-\alpha/2,n-1}\sqrt{\frac{s_x^2}{n}} \tag{4-7}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4-8}$$

where

$\bar{X}$ = the sample grand mean of X in the combustor subgroup
n = the sample size = the number of sites in the combustor subgroup
i = the index for the ith of n sites in the combustor subgroup.

[1] Comparisons based on hypothesis testing and on overlap of confidence intervals will generally result in the same conclusion even though they are not exactly equivalent.

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n-1} \tag{4-9}$$

where

s^2 = the sample estimate of the combustor subgroup variance of X
$t_{1\text{-}\alpha/2,n\text{-}1}$ = the (1-α/2)th quantile of the t-statistic associated with n-1 degrees of freedom.

Note: when $n \geq 60$, the (1-α/2)th quantile of the z-statistic may be substituted for $t_{1\text{-}\alpha/2,n\text{-}1}$; e.g., for a 95 percent confidence interval, use z=1.96.

4.2.3 Point Estimate of the pth Population Percentile (X_p) of X

Assume that we have a random sample of n measurements of the continuous variable X and that the sample values are ordered from the smallest to the largest and are denoted as $x_{i,}$ i=1,2,…,n, where i=the rank of the ith ordered value of X. Let p=i/100 = the proportion of the sample values with $X \leq x_i$ and define np=j+g, where j is the integer part and g is the fractional part of the quantity, *np*.

Then the sample estimate of the pth population percentile of X, X(p), is computed as follows (SAS, 1990):

$$\hat{X}_{(p)} = \begin{cases} \left(x_i + x_{i+1} \right) / 2 \;\; if \;\; g = 0 \\ x_i + 1 \;\; if \;\; g > 0 \end{cases} \tag{4-10}$$

Note: This estimator is distribution-free; i.e., it is robust to the effects of the actual shape of the underlying population distribution. Parametric estimators (e.g., lognormal percentile estimators; Gilbert, 1987) will be unbiased only if the specified population distribution actually fits the observed data. Thus, prior to computing the parametric percentile estimators, goodness-of-fit tests (e.g., Shapiro-Wilk's test) should be used to confirm that the sample data could have come from the specified parametric distribution. When sample sizes are small (*n*<20), goodness-of-fit tests tend to accept the null hypothesis that the data fit the parametric distribution, regardless of the actual goodness of fit to the parametric model. Moreover, in cases where the parametric assumptions hold, the nonparametric estimates of the percentiles have been shown to be very close to the parametric estimators. However, when the parametric assumptions fail, the nonparametric estimators provide truer estimates of the population percentiles. Because of their robustness in small sample situations, nonparametric, rather than parametric, percentile estimators were employed throughout this report.

4.2.4 Interval Estimate of the pth Population Percentile (X_p) of X

Given a random sample of *n* measurements of the continuous variable X, if the sample values are ordered from the smallest to the largest, the upper and lower 100×1-α% confidence bounds on the pth population percentile, X_p are the values of X that have ranks *r* and *s*, respectively, where *r* and *s* are calculated as follows (Altman et al., 2000):

$$r = round\left[np - \left(z_{1-\alpha/2} \sqrt{np\left(1-p\right)} \right) \right] \tag{4-11}$$

$$s = round\left[1 + np + \left(z_{1-\alpha/2} \sqrt{np\left(1-p\right)} \right) \right] \tag{4-12}$$

where

n = the sample size
p = the percentile proportion (e.g., p=0.90 for the 90th percentile)
round $\Rightarrow$ round to the nearest integer.

Note: this estimator is distribution free; i.e., it is valid for any population distribution.

4.3 Hypothesis Tests

4.3.1 Overview

This section describes several statistical tests of hypotheses that were used to compare attributes of the distributions of chemical-specific stack emission variables and variables related to stack characteristics and ambient weather for stacks from Phase I and Phase II combustors. Tests were performed for the equality of the correlations between pairs of selected variables and for equality of Phase I and II distributional parameters. The parameters included the mean, median (50th percentile), variance, and, as permitted by data limitations, the 10th, 25th, 75th, and 90th percentiles. As described in Section 4.2, "X" denotes a variable that is being compared between a Phase I subgroup (e.g., incinerators) and a Phase II subgroup (e.g., liquid boilers) and it is assumed that the parameter estimates are based on *m* Phase I combustor sites and *n* Phase II combustor sites. The primary hypotheses of interest for each variable X are described along with the associated tests. Additional details can be found in the references.

Table 4-1 lists the population parameters that were compared. "Population" can be interpreted as the entire set of combustors in some specified combustor subgroup of Phase I (e.g, incinerators) or Phase II (e.g., liquid boilers), that were operating in the United States during the period of the study. "X" denotes a variable (e.g., stack height) that is being compared between Phase I and II data; the notation X_1 and X_2 will be used to explicitly reference, respectively, Phase I and Phase II subgroup measurements and estimates. More generally, the notation X_i will be used when referring to aggregate measurements or estimates that can come from either a Phase I or Phase II subgroup. This notation is used throughout this section of the report. As indicated in the table, all parameters will be estimated separately within each phase.

Table 4-1. Population Parameters Compared

Population Parameter and Notation			Estimate of Parameter and Notation		
Name	**Phase 1**	**Phase 2**	**Name**	**Phase 1**	**Phase 2**
Population Size	M	N	Sample Size	m	n
Population Mean	μ_1	μ_2	Sample Mean	$\bar{X}_1$	$\bar{X}_2$
Population Standard Deviation	σ_1	σ_2	Sample Standard Deviation	s_1	s_2
Proportion of Population with X#C, where C = given constant	P_1	P_2	Sample Proportion of Observations with $X \leq C$	p_1	p_2
Correlation of Two Variables	ρ_1	ρ_2	Spearman Correlation	r_1	r_2
Population Percentile	$X_{1(p)}$	$X_{2(p)}$	Value of the variable X such that X is $\geq$ p percent of all other values of X in the sample	$x_{1(p)}$	$x_{2(p)}$

4.3.2 Test for Equal Variances

The hypothesis of equal variances for two populations for a given variable X is formally stated as follows: H_0: $\sigma_1 = \sigma_2$ vs. H_A: $\sigma_1 \neq \sigma_2$. A test for equality of the variances of X was conducted by using an F test. The statistic

$$F = \text{maximum variance/minimum variance} = \max\left(s_1^2, s_2^2\right)/\min\left(s_1^2, s_2^2\right) \qquad (4\text{-}13)$$

is compared to the tabulated F distribution. If a Type I error rate of 0.10 is desired, F is compared to the 95th percentile of the F distribution with m-1 and n-1 degrees of freedom if $s_1 > s_2$ and to the 95th percentile of the F distribution with n-1 and m-1 degrees of freedom if $s_1 < s_2$.

4.3.3 Test for a Common Median

The hypothesis for comparing population medians is stated as H_0: $X_1(50) = X_2(50)$ vs. H_A: $X_1(50) \neq X_2(50)$.

The generalized Wilcoxon rank sum statistic (Conover, 1999) can be used if the Phase I and II data are non-normal but have similar variance. This form of the Wilcoxon test is generalized in the sense that it allows for tied values of X within phases. The test statistic is computed as

$$T = \frac{\left[\sum_{i=1}^{n} R\left(X_i\right)\right] - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N\left(N-1\right)}\sum_{i=1}^{N} R_i^2 - \frac{nm\left(N+1\right)^2}{4\left(N-1\right)}}} \qquad (4\text{-}14)$$

where

$R(X_i)$ = the rank of X_i in the pooled-over-phases sample (ties are replaced with the average of the ranks)
m = the Phase I sample size
n = the Phase II sample size
N = n+m.

Note that the summation in the numerator is over only the ranks associated with the Phase II X values. This test evaluates whether the distributions of the Phase I and Phase II values of X_i differ from one another *only* on the basis of an unspecified shift parameter θ, such that corresponding quantiles (e.g., the medians) differ from one another by a constant value, θ. The null hypothesis evaluated by the Wilcoxon rank sum test is that θ=0, implying that the distributions of the Phase I and Phase II values of X are approximately identical, and hence that their medians are also equal. The alternative hypothesis is that the two distributions are shifted θ units apart. This test is valid if (1) the distributions are approximately symmetric, and (2) they have approximately the same variances. Thus, if the distributions are actually lognormal and/or if the variances are unequal, the results of the Wilcoxon test may be incorrect.

4.3.4 Test for a Common Generalized Percentile

To test whether Phase 1 and 2 share a given population percentile (median or other), the Phase 1 and Phase 2 data were combined and C, the combined estimated percentile, was determined. The test is formulated as a test of the respective proportions, P_1 and P_2, where P_i is the proportion of the Phase i population with levels of X less than C: H_0: $P_1 = P_2$ vs. H_A: $P_1 \neq P_2$. This test was carried out via a chi-square test, and is appropriate if both phases have reasonably large sample sizes. After determining C, the number of observations below and above C for each phase were determined and the table of counts shown in Table 4-2 was constructed.

Table 4-2. Counts for Common Percentile Test

Population	No. Observations with X≤C	No. Observations with X>C	Total
Phase 1	m_1	m_2	m
Phase 2	n_1	n_2	n
Total	n_1+m_1	n_2+m_2	n+m

Note that the estimates of P_1 and P_2 are $p_1 = m_1/m$ and $p_2 = n_1/n$, respectively. The chi-square statistic,

$$\chi^2 = \frac{\left[n_1 m_2 - n_2 m_1\right]^2 \left(n+m\right)}{nm\left(n_1+m_1\right)\left(n_2+m_2\right)} \qquad (4\text{-}15)$$

is compared to the percentiles of a chi-square distribution with 1 degree of freedom. The output from the test includes a warning if any of the counts in the table are less than five. This can occur (1) if there are extreme differences in the Phase 1 and 2 distributions, or (2) if the sample size for one or both Phases is small. In case (1), the chi-square test will likely be statistically significant. In case (2), it will generally not be significant Sand the warning should be regarded as an indication that the test is based on inadequate data. There is obviously a greater likelihood of case (2) occurring if more extreme percentiles are chosen.

4.3.5 Test for Common Correlations Between Two Variates

An approximate test for testing whether the correlations between two variables (denoted as X and Y) are the same for Phases 1 and 2 (i.e., a test of H_0: $\rho_1 = \rho_2$ vs. H_A: $\rho_1 \neq \rho_2$) is available for large sample sizes (*m*, *n*>20). Spearman (rank) correlations, r_1 and r_2, were first estimated for each phase. (Spearman correlations were used since they are not sensitive to outliers and do not depend on the scale of measurement.) These estimated correlations for each phase were then transformed, as follows:

$$z_1 = 0.5\ln\left[\frac{1+r_1}{1-r_1}\right] \quad \textit{and} \quad z_2 = 0.5\ln\left[\frac{1+r_2}{1-r_2}\right] \qquad (4\text{-}16, 4\text{-}17)$$

$$\rho = \frac{\sum_{j=1}^{n} R\left(X_j\right)R\left(Y_j\right)T}{\sqrt{\left(\sum_{j=1}^{n} R^2\left(X_j\right) - T\right)\left(\sum_{j=1}^{n} R^2\left(Y_j\right) - T\right)}} \qquad (4\text{-}18)$$

$$T = n\left(\frac{n+1}{2}\right)^2 \qquad (4\text{-}19)$$

where

ρ = Spearman's Corr. Coef.
$R(X_j)$ = the rank of X (or mean rank of tied X values) measured at the jth combustor site
$R(Y_j)$ = the rank of Y (or mean rank of tied Y values) measured at the jth combustor site
n = the number of of sites in the Phase-i combustor subgroup at which both X and Y were measured.

Then the test statistic

$$d = \frac{z_1 - z_2}{\sqrt{\frac{1}{m-3} + \frac{1}{n-3}}} \tag{4-20}$$

was computed. For a Type I error rate of approximately 0.10, the absolute value of *d* would be compared to the 95th percentile of the standard normal distribution.

4.4 Cumulative Distribution Function (CDF) Plots

A quantile is a statistical quantity that provides a measure of relative standing of a given observed value with respect to other observations. If *x* is the p^{th} quantile for a variable X, then at least 100p percent of the values in the data set lie at or below *x*, and at least 100(1-p) percent of the values lie at or above *x*. For example, the 0.95 quantile has the property that 95 percent of the observations lie at or below *x* and 5 percent of the data lie at or above *x*. A quantile plot for a variable X is a graphical representation of the data in which the vertical axis represents the observed values of X and the horizontal axis gives quantitative values from 0.0 to 1.0, with each point plotted according to the fraction of the points that it exceeds—that is, that point's associated quantile. Assume that X_1, X_2, ..., X_n represent the *n* observed values of a given variable arranged in ascending order. For each i, compute the fraction,

$$f_i = \frac{i - 0.5}{n} \tag{4-21}$$

Because quantile plots are plots of the cumulative probability of observing a value of X that is less than or equal to some value, X_i, they are often called cumulative distribution plots (or CDF plots). By convention, CDF plots interchange the axes of the quantile plots; i.e., a CDF plot is a quantile plot that has the cumulative probability on the vertical axis and the corresponding observed values of X on the horizontal axis. The CDF plot is a plot of the pairs (f_i, X_i), with straight lines connecting consecutive points. A CDF plot can be used to read quantile information such as the median and quartiles. This can be facilitated by drawing horizontal lines at the 0.25, 0.50, and 0.75 points on the vertical axis to mark the quartiles and median values (or any other quantiles of interest) and then noting where these lines intersect the plotted line that connects the pairs. In addition, the plot can be used to determine the density of the data points. For example, are all the data values close to the center with relatively few values in the tails or are there a large number of values in one tail with the rest evenly distributed? The density of the data is displayed through the slope of the graph. A large number of data values has a steep slope, i.e., the graph rises quickly. A small number of data values has a small slope, i.e., the graph is relatively flat. Thus, one can determine whether the data are relatively uniformly distributed or whether there are large clusters of points. A CDF plot can also be used to determine if the data are skewed or symmetric. A CDF plot of data that are skewed to the right will appear flatter at the top right than the bottom left, whereas a CDF plot of data that are skewed to the left will appear to be flatter near the bottom left of the graph and then become steeper. If the data are symmetric then the top portion of the graph will stretch to the upper right corner in the same way that the bottom portion of the graph stretches to the lower left, creating an S-shape.

Several CDF plots summarizing and comparing the distributions of estimated regression coefficients in various populations of combustor sites are presented in Appendix B. Pairs of CDFs that overlay or nearly overlay each other suggest that the underlying population distributions are very similar and perhaps even identical. The Wilcoxon rank sum test actually evaluates the null hypothesis that the distance between a pair of CDFs is zero. When this is true, the CDFs will overlay each other and corresponding percentiles and the means of the two populations are not significantly different from one another.

4.5 Interpretation and Limitations Due to Sample Size

4.5.1 Overview

The principal objective of the statistical analyses was to obtain evidence in support of the hypothesis that associations between measured risk estimates, emissions of various kinds, combustor characteristics, and meteorological variables that were actually calculated from the data obtained from the original Phase I combustors were predictive of the corresponding associations and risks in the Phase II combustors or in the new Phase I combustors for which no actual risk estimates were available. As part of the weight-of-evidence approach to this problem, statistical hypothesis tests were carried out and confidence interval estimates were computed from the data. In general, two types of tests were run. So-called two-sample tests (i.e., F-tests, t-tests, z-tests, and Wilcoxon rank sum tests) evaluated the null hypothesis that some parameter (e.g., the population mean TEQ emission rate) measured in the Phase I population was not significantly different from the corresponding parameter in the Phase II or New Phase I population. The alternative hypothesis was always two-sided—that is, that the corresponding parameters were different (as opposed to larger or smaller). The second type of test was the Shapiro-Wilk’s test of normality (or lognormality) wherein the null hypothesis was that the data were (log)normally distributed vs. the alternative that the data were not (log)normally distributed. These tests were used to determine the appropriateness of a log-transformation and to examine the reasonableness of the normality assumption underlying several of the two-sample statistical tests. In this section, a number of limitations and caveats regarding the interpretation of these tests and the confidence interval estimates will be discussed.

4.5.2 Sample Size Issues

Table 4-3 summarizes the various statistical tests and confidence interval estimates that were used to compare the population parameter estimates for the two groups of combustors. Two numbers are entered into the body of the table for every statistical test that was used to test the null hypothesis of equal group parameters (except the chi-square test). The first number is the minimum *combined* sample size of the two groups that is recommended for the validity of the associated statistical test. The second number is the recommended minimum sample size of the *smaller* of the two groups. For the chi-squared test (Section 4.3.4), the minimum number of combustor sites in *each cell* determines the validity of the test. The recommended sample size for the Shapiro-Wilk’s test (not entered in Table 4-3) is 11 combustors per group; it is not recommended that one should attempt inferences on the form of the population distribution (e.g., the distribution of TEQ emissions from solid boilers) from samples of less than 11 combustor sites.

All of the confidence interval formulae in Section 4.2 are functions of the standard normal deviate (i.e., z) and as such, are usually referred to as normal approximations. Although most statistical texts recommend sample sizes of at least 30 for normal approximations, that recommendation is relaxed here to a minimum of 20.

In general, larger sample sizes produce more valid and reliable estimates and statistical tests because

- On average, larger samples tend to be more representative of the population variability than are smaller samples
- The statistical power of the test (i.e., the likelihood of rejecting the null hypothesis) tends to be unacceptably low for small samples (e.g., $n<20$).

The issue of representativeness is most easily understood by considering the estimate of the population variance. The variance is the average squared deviation of the population members from the population mean. The best estimate of the variance should therefore be based on a sample(s) that includes some of the most extreme values (i.e., tail values) of the population distribution. The larger the sample, the more likely it is to include such relatively uncommon values; conversely the smaller the sample, the less likely it is to include extreme values. Thus, if small samples (e.g., $n=8$) are repeatedly drawn from a parent population that has a large variance, the majority of the samples are likely to seriously underestimate the population variance, while an occasional sample will grossly overestimate it. This situation is illustrated by the shot pattern shown in panel B of Figure 4-1; very few of the estimates are in the bull’s-eye area. Hence, any statistical tests (e.g., the F-test for equal variances) based on small sample estimates of the variance have a high probability of giving incorrect results. This problem applies to all of the population parameter estimates in Table 4-1.

Table 4-3. Recommended Minimum Samples Sizes for Two-Sample Test Statistics and One-Sample Confidence Interval Estimators

Estimated Population Parameter	F-test	T-test	WRS Test	Z-test	ChiSq Test	Normal Approx. C.I.[a]
Mean		30,10[b]	20,10[b]			20/group
Variance	30,10[b]					20/group
Spearman Rho				30,10[b]		
Percentile						
10th					5/cell[c]	20/group
25th					5/cell[c]	20/group
50th					5/cell[c]	20/group
75th					5/cell[c]	20/group
95th					5/cell[c]	20/group

[a] Table entries are the minimum sample sizes for each group (e.g., Phase 1 and Phase 2)

[b] Table entries are minimum *combined* Phase I and II group sample size, minimum required sample size for the *smaller* of the Phase I and II groups

[c] Minimum sample size per cell of 2×2 table (Section 4.3.4)

As discussed in Section 4.1.1, a Type II error occurs when the null hypothesis is erroneously accepted. In the case of the two-sample tests summarized in Table 4-2, this amounts to erroneously deciding that (for example) the Phase I and Phase II population parameters are *not* significantly different. Clearly, the consequences of this type of error in a given study may be as serious as, or more serious, than a Type I error. Hence, extreme caution is recommended in interpreting any nonsignificant (e.g., $p>0.05$) results in which either the combined or the individual group minimum sample size requirements (Table 4-1) are not met.

Type II error rates are usually expressed in terms of statistical power, where

$$\text{Power} = 1\text{-Prob(Type II error)}. \qquad (4\text{-}22)$$

That is, power is the complement of the Type II error rate. In terms of the judicial analogy used earlier, power is the probability of *correctly* convicting a guilty person. Thus we desire that our statistical tests have high power so as giving us high confidence that we have *correctly* identified all of the pairs of Phase I and II combustor subgroups whose population parameters are different from one another. How one defines “high” is subjective. However, if we consider that a coin-flip decision of guilt or innocence has a power of 0.50, it is obvious that we desire the power of a test to be greater than 0.50. In many experimental studies, the minimum acceptable power is set at 0.80. However, in a study such as this one, where potentially hazardous outcomes are being considered, a higher standard may be required.

The power of a two-sample test is a function of four factors: (1) the variance of the two samples, (2) the two sample sizes, (3) the prespecified α-level (i.e., the maximum allowable Type I error rate), and (4) the minimally important difference between the two populations. Defining the minimally important difference is problematic for the mean and the variance;

however, because both proportions and correlation coefficients are constrained to take on values (respectively) from 0 to 1 and from -1 to +1, one can easily compute power estimates in (say) increments of 0.10 over the entire range of the expected parameter values. This has been done in Tables C-1 and C-2 of Appendix C. The tables provide power estimates for different combinations of sample size and minimally important differences, at several different α-levels.

The minimally important differences in Table C-1 are labeled "relative differences in percentiles." These values are actually the difference in the proportion of the measured group (Phase I) whose response values (e.g., TEQ emission rate) are below the pooled population P^{th} percentile (e.g., the 25th percentile) and the corresponding proportion in the group for which no risk was measured (Phase II or New Phase I combustors). Clearly the larger this difference is, the more likely it is that the null hypothesis of equal percentiles is wrong. The minimally important difference in correlation coefficients is labeled "DELTA" in Table C-2. In both Table C-1 and C-2, the desirable combinations of sample size, α-level, and minimal differences will be associated with powers > 0.80. Similarly, for a given pair of sample sizes, power and α-level, one can find the smallest observed difference that the statistical test will "recognize" as being statistically significant. For example, for a test of the equality of a Phase I correlation coefficient computed from a sample of 25 combustor sites vs. a correlation coefficient computed from a sample of 8 Phase II sites, such that the power is = 0.80, the two correlation coefficients will have to differ by more than 0.60. Thus for this combination, only very large differences will be "detectable." If differences in correlation as small as ± 0.25 are deemed to be potentially biologically important, then clearly estimates from sample sizes of 8 or less will be an inadequate basis for decisions.

4.5.3 Percentile Estimates

Percentile estimation poses special problems in small samples and populations. Literally, a percentile tells one what percentage of a population is less than a particular value. Because percentages are measured in increments of 1/100, it is difficult to obtain meaningful estimates of them from very small samples or for very small finite populations. For example, if we have a sample of 5 values of X = 7, 9, 12, 15, 29 and we try to estimate percentiles in the upper tail of the population distribution, we will find that all of the percentiles from 81 to 100 have the value 29, the sample maximum. That is, we cannot distinguish the 81st from the 99th percentile in this sample. Moreover, if a population is very small (e.g., N=5), it is hard to justify any attempts to partition it into in hundredths; for such small populations, percentiles are not useful or meaningful descriptive statistics.

Because of this problem and the sample size limitation of 5 combustor sites per cell of the 2×2 table for the χ^2 test for equal percentiles (Section 4.3.4), a system of contingencies has been developed for the comparison of percentile estimates from small samples (see Section 5, Table 5-7). The rationale for the system is that (1) tail percentiles are more difficult to estimate accurately from small samples, and (2) in small sample situations, comparison of confidence estimates for overlap is both more informative and more robust than hypothesis testing. For example, in comparing the 95 percent confidence interval estimates of two sample estimates of the 75th percentile, we might see values like this:

75th percentile of X in Phase I = 17 (1, 94)
75th percentile of X in Phase II = 77 (20, 194).

The point estimates suggest that the difference between the two 75th percentile estimates is actually quite large (i.e., 60 units) but because the confidence intervals are badly inflated due to the small sample sizes, they overlap considerably and, in effect, we are unable to detect a significant difference. Thus, inflated confidence intervals are the counterpart to low statistical power in statistical hypothesis tests that occurs when sample sizes are small; like those tests, comparison of such confidence intervals will often lead to Type II decision errors. Thus, pairs of confidence interval estimates, like the two above, provide a flag or warning of the apparent high risk of a Type II decision error; a warning that is particularly appropriate when the sample sizes are below those shown in Table 4-3.